Approximation Theory of Deep Neural Networks: Part 2

Research school: HiDaDeeL

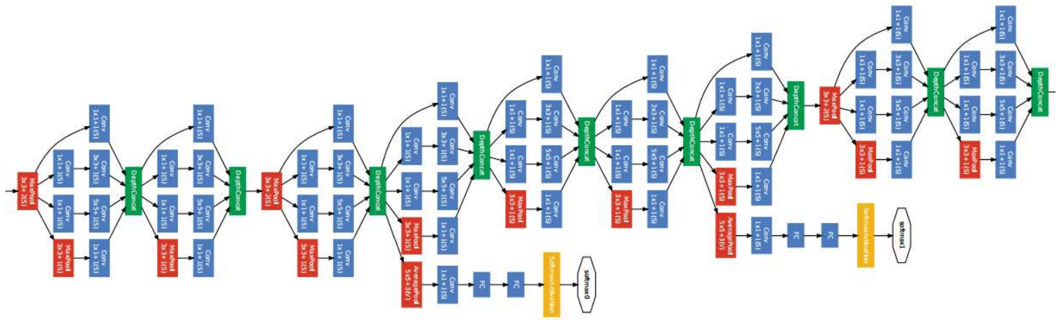
16.05 – 20.05. 2022

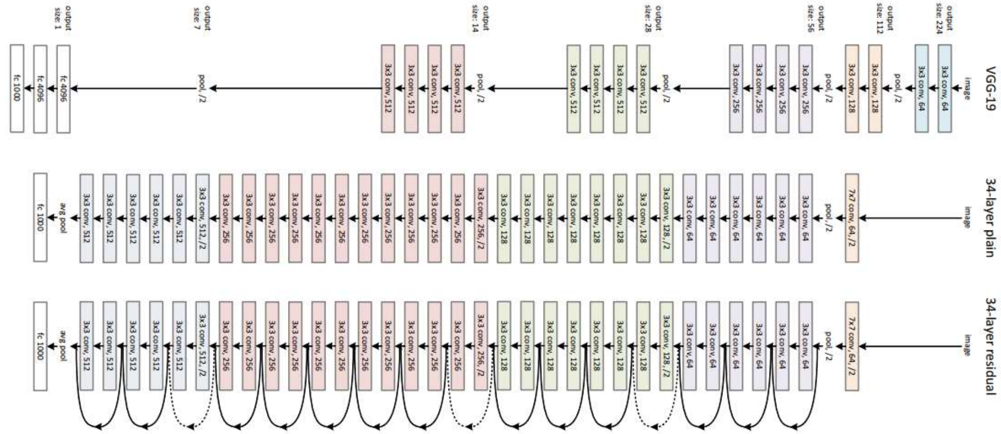
Nantes

Philipp Petersen

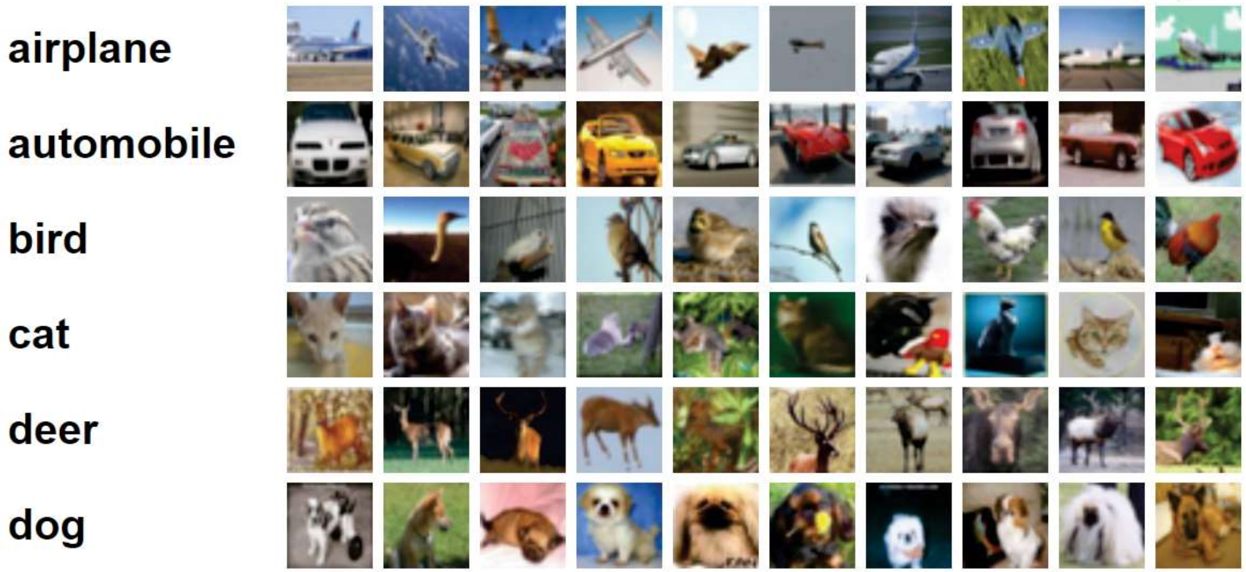
Topics:

- ## Deep vs shallow



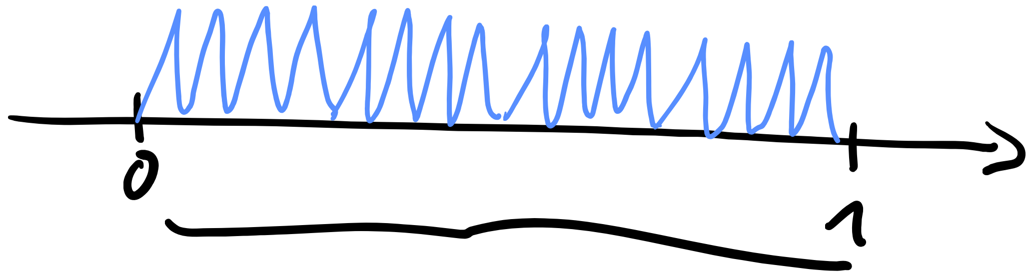


- ## Curse of dimension



Recall:

$$R(\Phi^1 \bullet \Phi^1 \bullet \dots \bullet \Phi^1) =$$



2^n hats.

On the other hand, shallow NNs generate at most $O(N)$ pieces.

Can we have a more precise trade-off?

Thm: Let $L \in \mathbb{N}$. Let ρ be a piecewise affine function with p pieces.
Then, for every NN Φ with $d = 1$, $N_L = 1$ and $N_1, \dots, N_L \leq N$, we have that $R(\Phi)$ has at most $(pN)^{L-1}$ affine pieces.

Proof:

Induction over L.

$L=2$ ✓

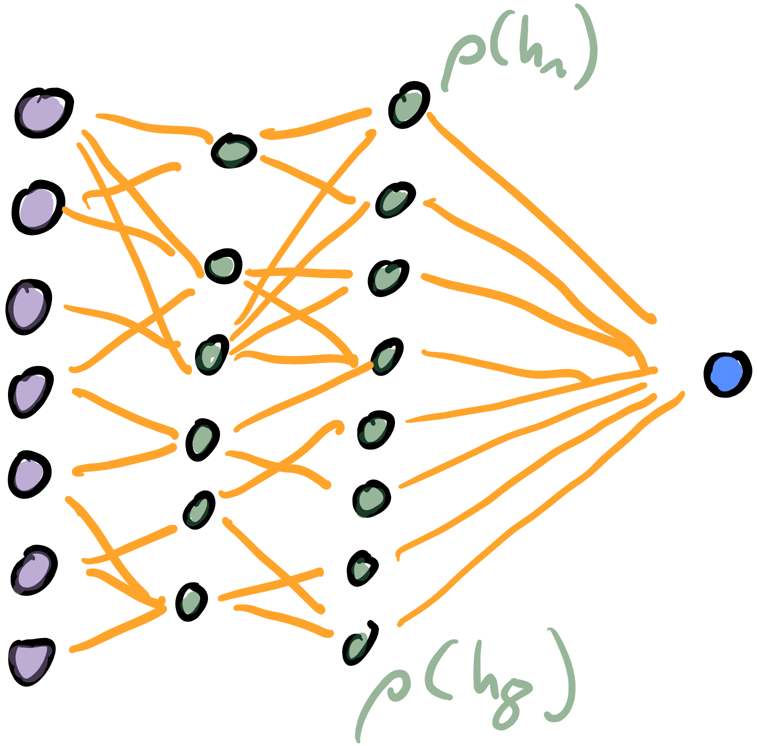
Note:
- f_1, f_2 pw. affine with n_1, n_2 pieces
  $\Rightarrow$ f_1+f_2 pw. affine with n_1+n_2 pieces.

- If f has n_1 pieces, then $\rho \circ f$ has at most $p \cdot n_1$ pieces.

Let Φ_{L+1} be a NN with $L+1$ layers.

$\Longrightarrow$

$$R(\Phi_{L+1}) = A_{L+1}\left[\rho(h_1(x)), \rho(h_2(x)) \dots\right] + b_{L+1}.$$

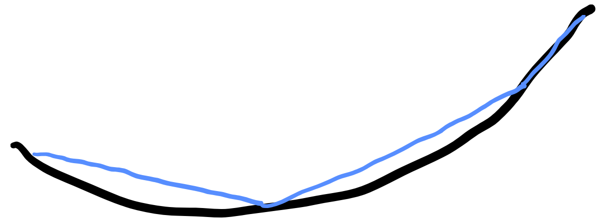


$$= \sum_{j=1}^{N_L} (A_{L+1})_j \, \rho(h_j(x)) + b_{L+1}$$

Realisations of NNs with L layers.

Cor. Let $L, d \in \mathbb{N}$. Let ρ be p.w. affine with p pieces. Then for every NN Φ with $N_L = 1$ and $N_1, \dots, N_L \leq N$, we have that $R(\Phi)$ has at most $(Np)^{L-1}$ pieces along every line.

Do we need many pieces?



Yes!

Prop: [Frenzen, Sasao, Butler; 2010]

Let $f \in \mathcal{C}^2([a,b])$, for $a < b < \infty$ so that f is not affine. Then there ex. $C(f) > 0$ s.t. for every $p \in \mathbb{N}$

$$\|g - f\|_\infty > c\, p^{-2}$$

for all g which are p.w. affine with p pieces.

Thm: [Eldan, Shamir; 2016, Yarotsky 2017]

Let $f \in C^2([0,1]^d)$ non-affine. Let $\rho: \mathbb{R} \to \mathbb{R}$ be p.w. affine with p pieces.

Then $\exists\, c(f) > 0$.

$$\| f - R(\Phi) \|_\infty \geq c(f) \left(p\, N(\Phi) \right)^{-2(L(\Phi)-1)} .$$

Conclusion:

- shallow NNs can, at best, achieve polynomial approx rates for smooth functions.
- Trade-off between depth and width.
- Extension to L^p is possible.
- What about non-ReLU activations?

Rotation invariant functions

Consider a function $f(x) = g(\|x\|_2^2)$.

If g is smooth, e.g. C^k, then we expect that f can be approximated with a NN with $\mathcal{O}(\varepsilon^{-\frac{1}{k}})$ weights up to error ε, with a deep NN. The constants depend linearly on the dimension d.

Reason : • $x \mapsto \|x\|_2^2 = \sum_{i=1}^{d} |x_i|^2$ is a sum of $1d$ squares.

• g is one dimensional.

What about shallow nets?

Thm: [Eldan, Shamir; 2016]

Let ρ be the ReLU. There ex. constants $c, C > 0$ s.th: For every $d \in \mathbb{N}$ with $d > C$ there ex. $g: \mathbb{R} \to \mathbb{R}$,

For all $n \leq c \cdot e^{cd}$:

$$\inf_{\substack{\Phi \text{ with arch} \\ (1,n,1)}} \| R(\Phi) - g(\|\cdot\|_2^2) \|_{L^2(K)} \geq c.$$

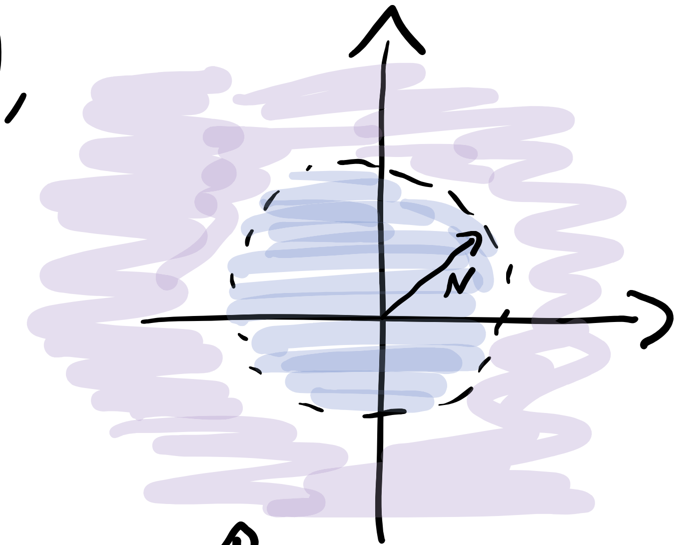
Proof:

$$\int g(\|x\|)\, e^{-2\pi i \langle x, \xi \rangle}\, dx$$

$$\cong \iint g(r)\, e^{-2\pi i r \langle \theta, \xi \rangle}\, r^{d-1}\, d\theta\, dr$$

$$\cong \int_{\theta \in S^{d-1}} \hat{g}^{(d-1)}(\langle \theta, \xi \rangle)\, d\theta \approx \tilde{g}(\|\xi\|_2),$$

where $\tilde{g}(r) = \int_{\theta \in S^{d-1}} \hat{g}^{(d-1)}(r\theta_1)\, d\theta.$



$\sim N^{-k}$ of mass outside.

We can choose g s.th. $\tilde{g}$ does not decay rapidly.

$$\Big\| \sum_{i=1} a_i^2 \rho(\langle a_i^1, \bullet \rangle + b_i^1) + b^2 - g(\|\bullet\|) \Big\|_{L^2(K)}$$

cut-off.

$$\gtrsim \Big\| \Big(\sum a_i^2 \rho(\langle a_i^1, \cdot \rangle + b_i^1) + b^2 - g(\|\cdot\|) \Big) \phi \Big\|_{L^2(\mathbb{R}^d)}$$

Plancherel: We can look at L^2 difference between

$$\sum_{i=1} \tilde{\mathcal{F}}\big(a_i^2 \rho(\langle a_i^1, \cdot \rangle + b^1) + b^2 \big) * \hat{\phi},$$

and $\tilde{g} * \hat{\phi}$.

It can be shown that

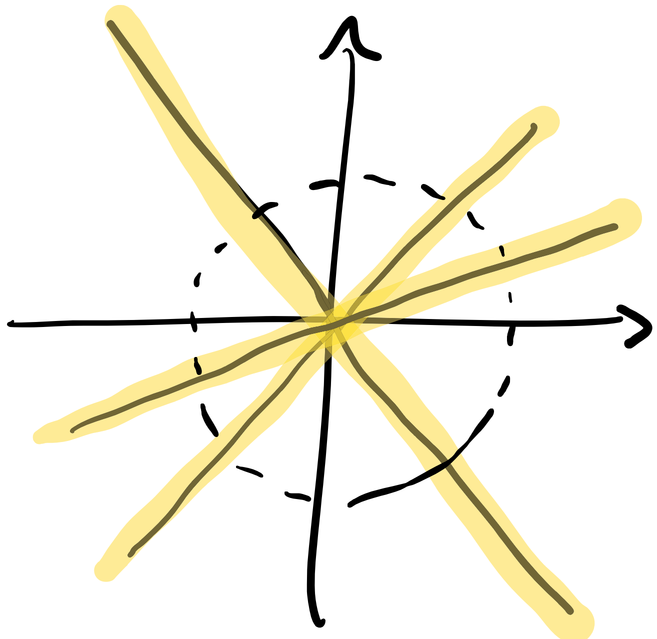
$$\mathcal{F}\left(\rho\left(\langle a_i^1, \cdot\rangle\right)\right) = \mathcal{F}\left(\rho \otimes \mathbb{1}_{\mathbb{R}^{d-1}} \circ R_{a_i}\right)$$

$$= \left(\mathcal{F}(\rho) \times \delta_{\mathbb{R}^{d-1}}\right) \circ R_{a_i}$$

is just supported on an arc along a_i.

$\Rightarrow$

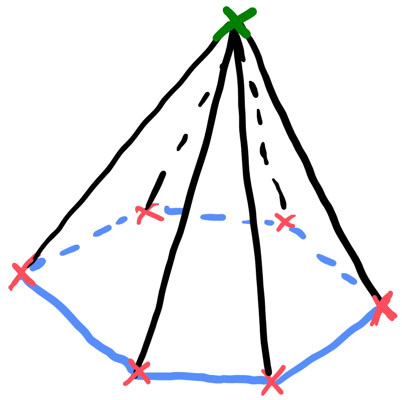
$$\mathcal{F}\left(\sum a_i^2 \rho\left(\langle a_i^1, \cdot\rangle + b_i^1\right) + b^2\right) * \hat{\phi} \approx$$



Volume of spheres $\approx$ r^d

Localised approximation

We have seen that one can build high-dimensional courant elements with deep ReLU NNs.



Thm: Let $d \geq 2$.

If $R(\phi)$ is compact for a ϕ with $L(\phi) = 2$, then $R(\phi) = 0$. (Activation function is ReLU.)

Proof:

$$R(\phi) = \sum_{i=1}^{m} a_i^2 \rho(\langle a_i^1, \cdot \rangle + b_i^1) + b^2$$

- If all a_i are different up to sign, then $R(\phi)$ has discontinuous derivatives on lines, except zero sets.

- if $a_i = \pm a_j \Rightarrow$ either cancellation $\Rightarrow$ we can remove terms from sum, or discontinuity remains, or sum is affine linear $\rightsquigarrow$ must be 0 to have cpt supp.

Curse of dimensionality

<u>Recall</u>: For $f \in C^k([0,1]^d)$:

NN Φ: $$\|f - R(\Phi)\|_\infty \leq M^{-\frac{k}{d}},$$
$$M(\Phi) \leq M.$$

[Yarotsky; 2017] $\rightsquigarrow$ This is optimal if $L \approx \log(M)$.

For image classification d = num of pixels.

$\rightsquigarrow$ This result does not explain why NNs work.

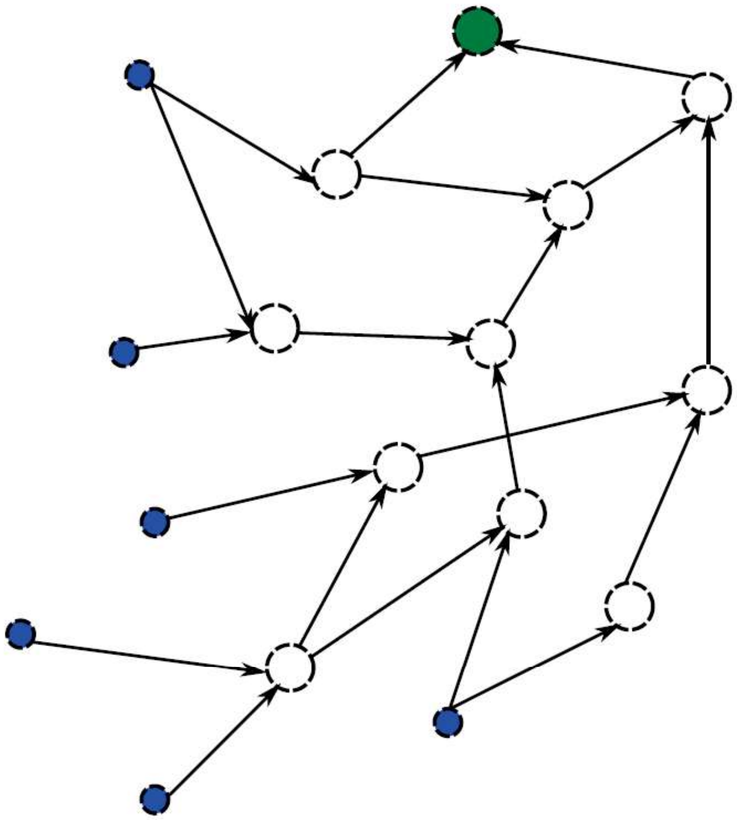
Compositional functions

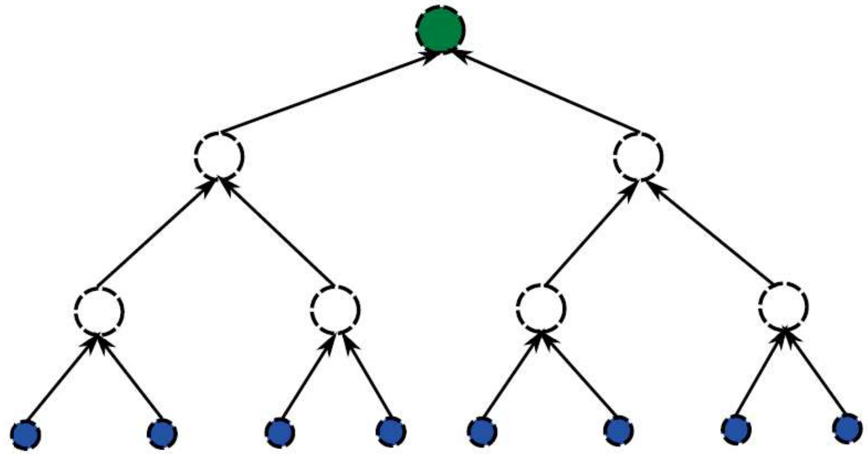
Not all high dimensional functions are problematic.

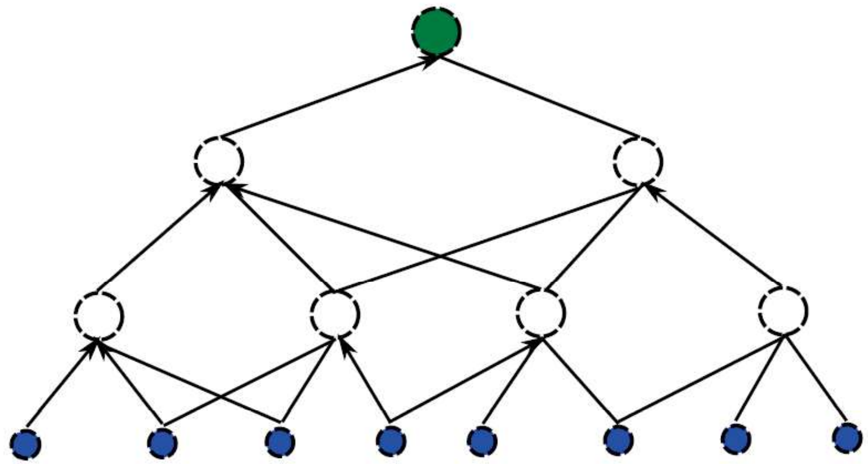
E.g. $x \longmapsto \|x\|_2^2 = \sum_{i=1}^{d} |x_i|^2$ is just a sum of d one dimensional functions and sums are simple.

$x \longmapsto \max\{x_i, i \in d\}$ can be found with d two-dimensional max operations.

Compositional functions [Mhaskar, Poggio; 2016]





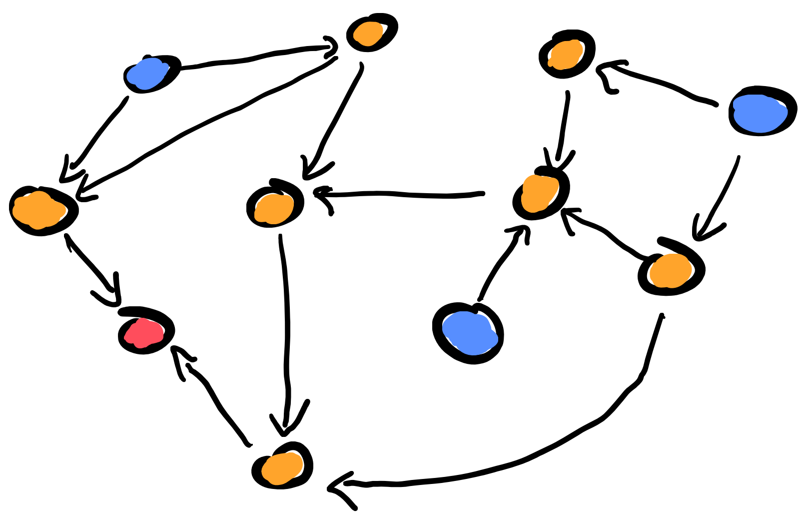


- Visual cortex
- Sensor networks

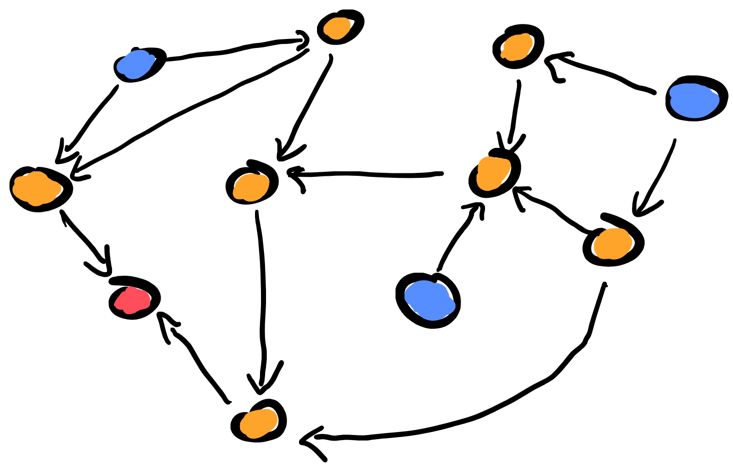
- Politics

Def:

Let $d, k, N \in \mathbb{N}$ and let $\mathcal{G}(d,k,N)$ be the set of directed acyclic graphs with N vertices, where the indegree of every vertex is at most k, the outdegree of all but one is at least 1 and the indegree of exactly d vertices is 0.



Def:



Let $C\mathcal{F}(d, k, N, s)$ be the set of functions f where we associate for $G \in \mathcal{G}(d, k, N)$ functions $f_{\eta_i} \in C^s$ with $\|f_{\eta_i}\|_{C^s} \leq 1$ to each vertex η_i and then compute f by computing the f_i's in the order of G.

Thm: Let $d, k, N, s \in \mathbb{N}$. Then, there ex. $C > 0$ s.th. for every $f \in C\mathcal{F}(d, k, N; s)$ and $\varepsilon \in (0, 1/2)$ there ex. a NN Φ_f with

$$L(\Phi_f) \leq C \cdot N^2 \log_2(k/\varepsilon)$$

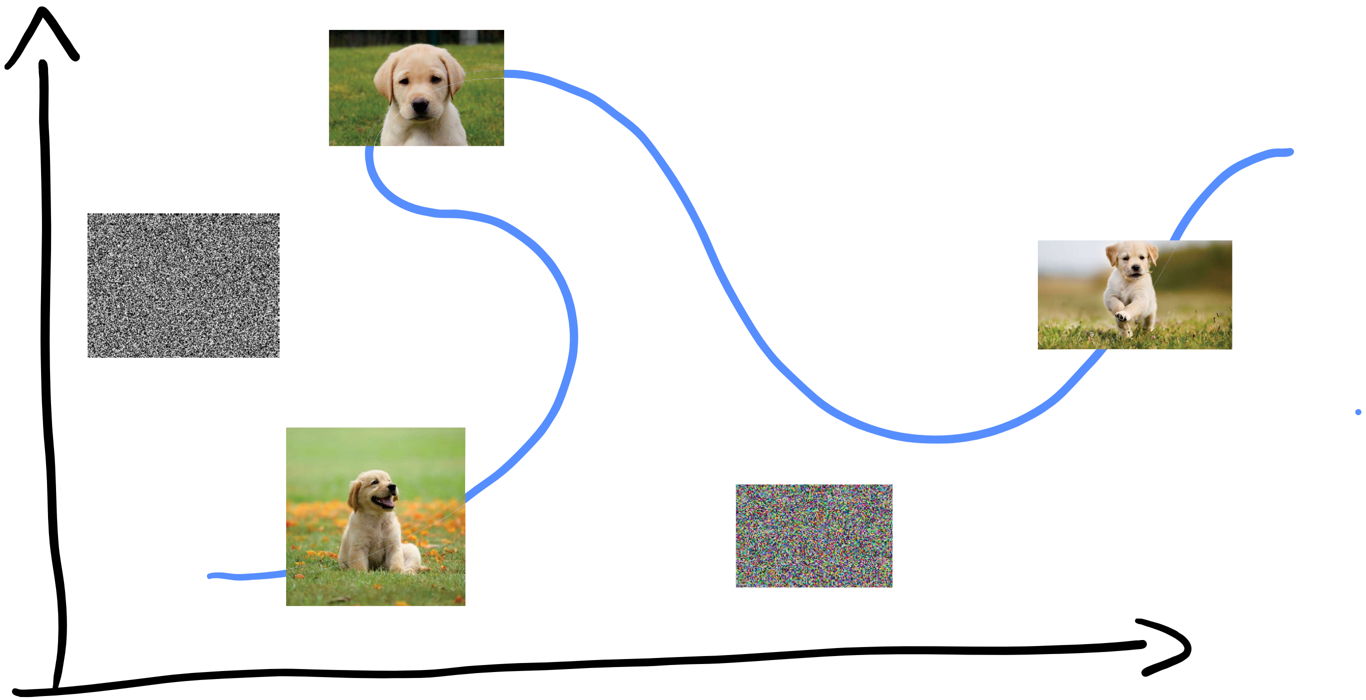
$$M(\Phi_f) \leq C N^4 (2k)^{\frac{kN}{s}} \varepsilon^{-k/s} \log_2(k/\varepsilon)$$

$$\|f - R(\Phi^f)\|_\infty \leq \varepsilon.$$

Activation function is ReLU.

The manifold assumption

Assumption: $\exists\ \Gamma \subset \mathbb{R}^D$, and Γ is a d dim manifold with $d \ll D$.
For $f: \mathbb{R}^D \to \mathbb{R}$, we only care about approx. on Γ.



Def: Let $\mathcal{M}$ be a smooth d-dimensional submanifold of $\mathbb{R}^D$. For $N \in \mathbb{N}$, $\delta > 0$, we say that $\mathcal{M}$ is (N,δ)-covered, if there ex. $x_1, \dots, x_N \in \mathcal{M}$ s.th.

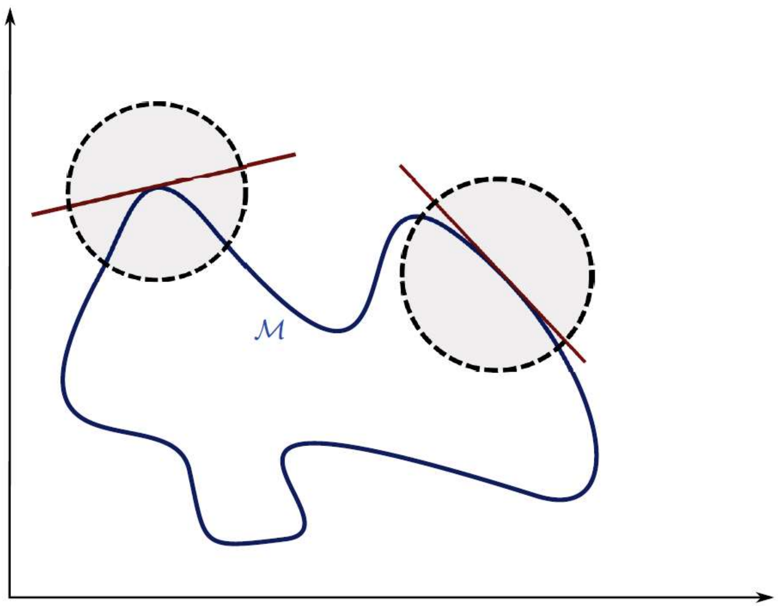
- $\bigcup_{i=1}^{N} B_{\delta/2}(x_i) \supset \mathcal{M}$,

tangent space

- the projection $P_i : \mathcal{M} \cap B_\delta(x_i) \longrightarrow T_{x_i}\mathcal{M}$ is injective, smooth, and

$$P_i^{-1} : P_i(\mathcal{M} \cap B_\delta(x_i)) \longrightarrow \mathcal{M}$$

is smooth.



Locally smooth functions

Def:

Let $\Gamma \subset \mathbb{R}^D$, be (N, δ)-covered.

Let $x_1, \dots, x_N$ be the centers of a cover.

For $f : \mathbb{R}^D \to \mathbb{R}$, we define

$$\| f \|_{C^k, \delta, N} = \sup_{i \in [N]} \| f \circ P_i^{-1} \|_{C^k} .$$

Let Γ be (N,δ)-covered, then there ex. a smooth partition of unity of Γ, subordinate to $(B_\delta(x_i))_{i=1}^N$. We call it $(\phi_i)_{i=1}^N$.

$$\Rightarrow \quad f(x) = \sum_{i=1}^N \phi_i \, f(x) = \sum_{i=1}^N \phi_i \cdot f \circ P_i^{-1}(P_i x)$$

$$= \sum_{i=1}^N \phi_i \cdot f_i (P_i x)$$

C^k on d-dim set

high-dim, but can be chosen

multiplication is OK

linear projection.

Thm: [Shaham, Cloninger, Coifman ; 2018][Chui, Mhaskar; 2018], ...

Let $D, k \in \mathbb{N}$, $\Gamma \subset \mathbb{R}^D$ be an (N, δ) covered d-dimensional manifold for $N \in \mathbb{N}$, $\delta > 0$. Then, there ex. $c > 0$, s.th. for every $\varepsilon > 0$ and $f \in C^k(\Gamma, \mathbb{R})$ with $\|f\|_{C^k, \delta, N} \leq 1$, there ex. a NN Φ s.th

$$\|f - R(\Phi)\|_\infty \leq \varepsilon,$$
$$M(\Phi) \leq c \cdot \left(\varepsilon^{-\frac{d}{k}} \log_2(1/\varepsilon)\right),$$
$$L(\Phi) \leq c \cdot \left(\log_2(1/\varepsilon)\right).$$

Here the activation function is the ReLU.

The Barron class

Definition: (Barron, 1993)

For $f = \int_{\mathbb{R}^d} \hat{f}(\xi)\, e^{i\langle \xi, \cdot \rangle}\, d\xi$, we define

$$C_f := \int_{\mathbb{R}^d} |\xi|\, |\hat{f}(\xi)|\, d\xi$$

and say $f \in \Gamma_C$, if $C_f \leq C$.

No curse of dimension

Proposition: (Barron; 1993/1992)

It holds that

$$\|f - f_n\|_{L^2(B(0),\mu)} \leq \frac{2C_f}{\sqrt{n}}$$

where f_n is a 2-layer NN* with n neurons.
(can be extended to L^∞ estimate.)

* under some assumptions on the activation function.

Some Barron functions

- $f \in \Gamma_C \Rightarrow \quad f(a \cdot - c) \in \Gamma_{|a| C}$
- $(f_i)_{i=1}^n \in \Gamma_C \Rightarrow \quad \sum_{i=1}^n \sigma_i f_i \in \Gamma_C$, if $\|\sigma_i\|_1 \leq 1$.
- General radial functions $\left(x \mapsto g(\|x\|_2^2)\right) \in \Gamma_{C_d}$, but $C_d \sim e^d$. (We have seen this before).
- Gaussian : $\left(x \mapsto e^{-\frac{\|x\|^2}{2}}\right) \in \Gamma_{C_d}$, where $C_d \leq \sqrt{d}$.

- Very smooth functions

$$\left\{ f \in W^{\lceil \frac{d}{2} + 2 \rceil, 2}, \ \|f\|_{W^{\lceil d/2 \rceil, 2}} \leq 1 \right\} \subseteq \Gamma_C .$$

Parametric problems

often high-dim. (discretised)

Problem: $\lambda \longmapsto f_\lambda$, where f_λ is the solution of some PDE, depending on λ.

Without using the structure of the problem this is a higly complex, high dimensional function.

Linear problem:

$$f_\lambda = A_\lambda^{-1} b_\lambda .$$

$\Longrightarrow$ $\lambda \mapsto f_\lambda$ consists of two steps. Building A_λ, b_λ and solving a linear system.

Assume: $\exists\, V_\varepsilon : \; f_\lambda \approx V_\varepsilon \underbrace{V_\varepsilon^T A_\lambda^{-1} V_\varepsilon}_{A^{-1}_{\lambda,\varepsilon} \in \mathbb{R}^{d(\varepsilon)\times d(\varepsilon)}} \underbrace{V^T b_\lambda}_{b_{\lambda,\varepsilon} \in \mathbb{R}^{d(\varepsilon)}}$

[$V_{\varepsilon,\lambda}$ would be possible too]

Thm: [Kutyniok, P., Raslan, Schneider; 2019]

$\exists\, \Phi : \quad M(\phi) \approx d(\varepsilon)^3$

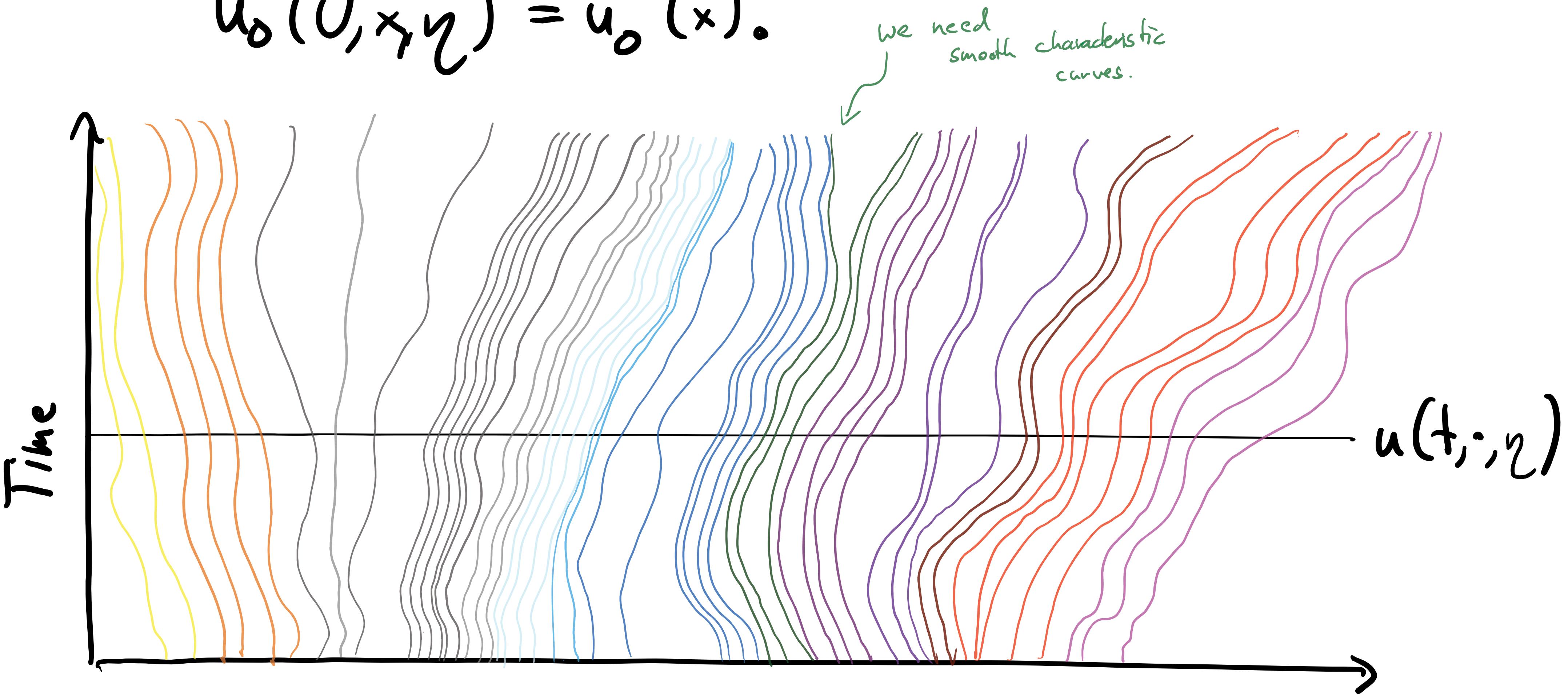
$R(\phi)\left(\left(A_{\lambda,\varepsilon}, b_{\lambda,\varepsilon}\right)\right) \approx f_\lambda$

$\lambda \mapsto f_\lambda$ can be approx. with NNs of size polynomial in $d(\varepsilon)$.

Parametric transport equations

$$\partial_t u(t,x,\eta) + V(t,x,\eta)\cdot\nabla_x u(t,x,\eta) = f(t,x,\eta)$$

$$u_0(0,x,\eta) = u_0(x).$$



Theorem: (Laakmann, P.; 2020)

Let $V \in C^k([0,T] \times \mathbb{R}^n \times [0,1]^D)$, $u_0 \in C^1$ approximable by NNs with rate r.

Then, for every $\varepsilon \in (0,1)$ and f sufficiently smooth a NN $\Phi^{u,\varepsilon}$ exists with $\|u - R(\Phi)\|_{L^\infty} < \varepsilon$ for u s.th.

$$\partial_t u(t,x,\eta) + V(t,x,\eta) \cdot \nabla_x u(t,x,\eta) = f(t,x,\eta)$$

$$u_0(0,x,\eta) = u_0(x).$$

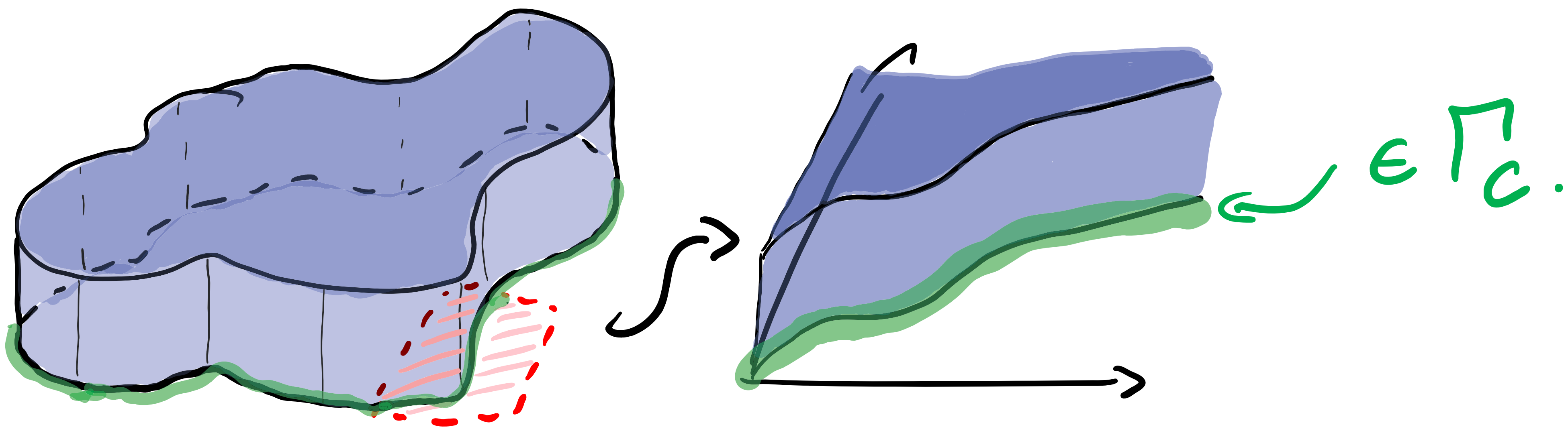
Here, $d = 1 + n + D$,

$$L(\Phi^{u,\varepsilon}) \lesssim \ln(1/\varepsilon)$$

$$W(\Phi^{u,\varepsilon}) \lesssim \varepsilon^{-1/r} + \varepsilon^{-\frac{d+1}{k-1}}.$$

Functions with Barron class singularities

$f = \chi_B$, where ∂B is locally in Γ_C.



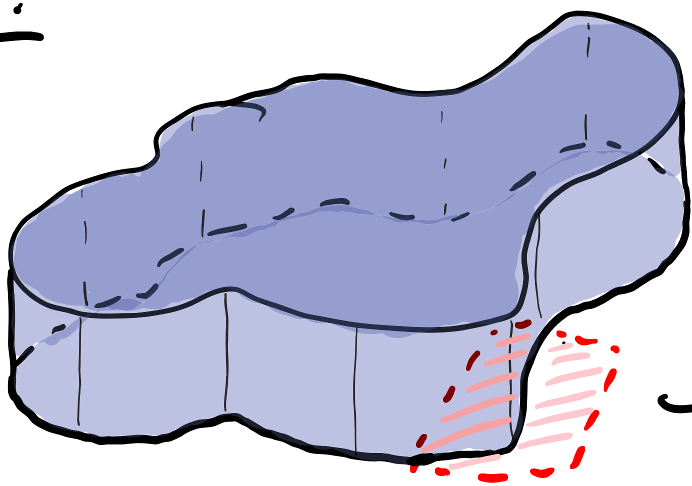
Theorem (Caragea, Petersen, Voigtlaender; 2020)

Let $f = \chi_B$ where B has Barron class boundary.
For every $N \in \mathbb{N}$, there exists a NN Φ with 4 layers and N neurons such that for each measure μ:

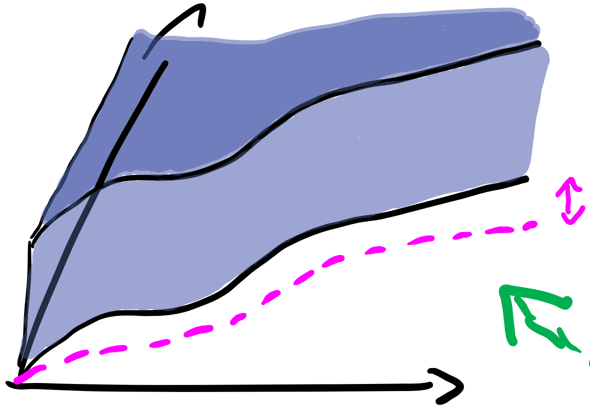
$$\mu\left(\left\{x \in \mathbb{R}^d : \chi_B(x) \neq R(\Phi)(x)\right\}\right) < C \cdot K \cdot d^{3/2} \cdot N^{-\alpha/2},$$

where C, α depend on μ, K on the size of a covering of ∂B. Also $0 \leq \Phi \leq 1$.

Proof:



a)



$$\chi_{\{x_1 \le \gamma(x_2,\dots,x_d)\}} = H(\gamma(x_2,\dots,x_d)-x_1)$$

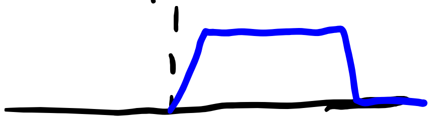
b)

Barron regular, hence

$$\Big\|\gamma - \sum_{i=1}^{N} \rho(\langle a_i,\cdot\rangle + b_i)\Big\|_\infty < 1/\sqrt{N}.$$

a)

$$\Lambda(x) = n\left[\rho(x) - \rho\left(x-\tfrac{1}{n}\right) - \rho\left(x - \tfrac{n-1}{n}\right) + \rho(x-1)\right]$$

$=$ 

$$\leadsto \rho\left(\sum_{i=1}^{d} \Lambda(x_i) - d + 1\right) \approx \chi_{[0,1]^d}(x).$$

Also, $\chi_\Omega(x+y-1) = y \cdot \chi_\Omega(x)$ for $y \in \{0,1\}$.

Conclusion

- Advantages of deep over shallow:

  Number of pieces (exponential in L), compactly supp. functions, radially symetric functions.

- Curse of dimension:

  Overcome in compositional functions, Manifold assumption, Barron class.

Bonus Round

(due to discussion yesterday)

Let $S \in \mathbb{N}^{L+1}$, then we denote by $NN(S)$ the set of NNs with architecture S.

Thm: [Raslan, P., Voigtlaender; 2020]
Let $\Omega \subset \mathbb{R}^d$ be compact and $S = (d, N_1, \dots, N_L) \in \mathbb{N}^{L+1}$ be a NN architecture. If $\rho: \mathbb{R} \to \mathbb{R}$ is cont., then

$$R: NN(S) \to L^\infty(\Omega)$$
$$\Phi \longmapsto R(\Phi)$$

is continuous. If ρ is locally Lipschitz, then R is locally Lipschitz.

Thm: Let $S = (N_0, N_1, \dots N_L) \in \mathbb{N}^{L+1}$ be a NN architecture, let $\Omega \subset \mathbb{R}^{N_0}$, and let $\rho: \mathbb{R} \to \mathbb{R}$ be Lipschitz continuous. Then, $R(NN(S))$ contains at most

$$\sum_{\ell=1}^{L} (N_{\ell-1} + 1) N_\ell$$

linearly independent centres.

Also $R(NN(S))$ is scaling invariant.

Cor.: Let $S=(N_0,\dots,N_L)\in\mathbb{N}^{L+1}$, let $\Omega\subset\mathbb{R}^{N_0}$ and let ρ be Lipschitz cont. If $R(NN(S))$ contains more than $\sum_{\ell=1}^{L}(N_{\ell-1}+1)\cdot N_\ell$ linearly indep. functions, then $R(NN(S))$ is not convex.

$\Longrightarrow$ $R(NN(S))+B_r(0)$ is only convex if it is dense. (if activation function facilitates universal approximation.)

$\longrightarrow$ relation to space filling curves.

Thm: For most activation functions ρ and for $S \in \mathbb{N}^{L+1}$

$$R(\mathcal{NN}(S)) \text{ is not closed}$$

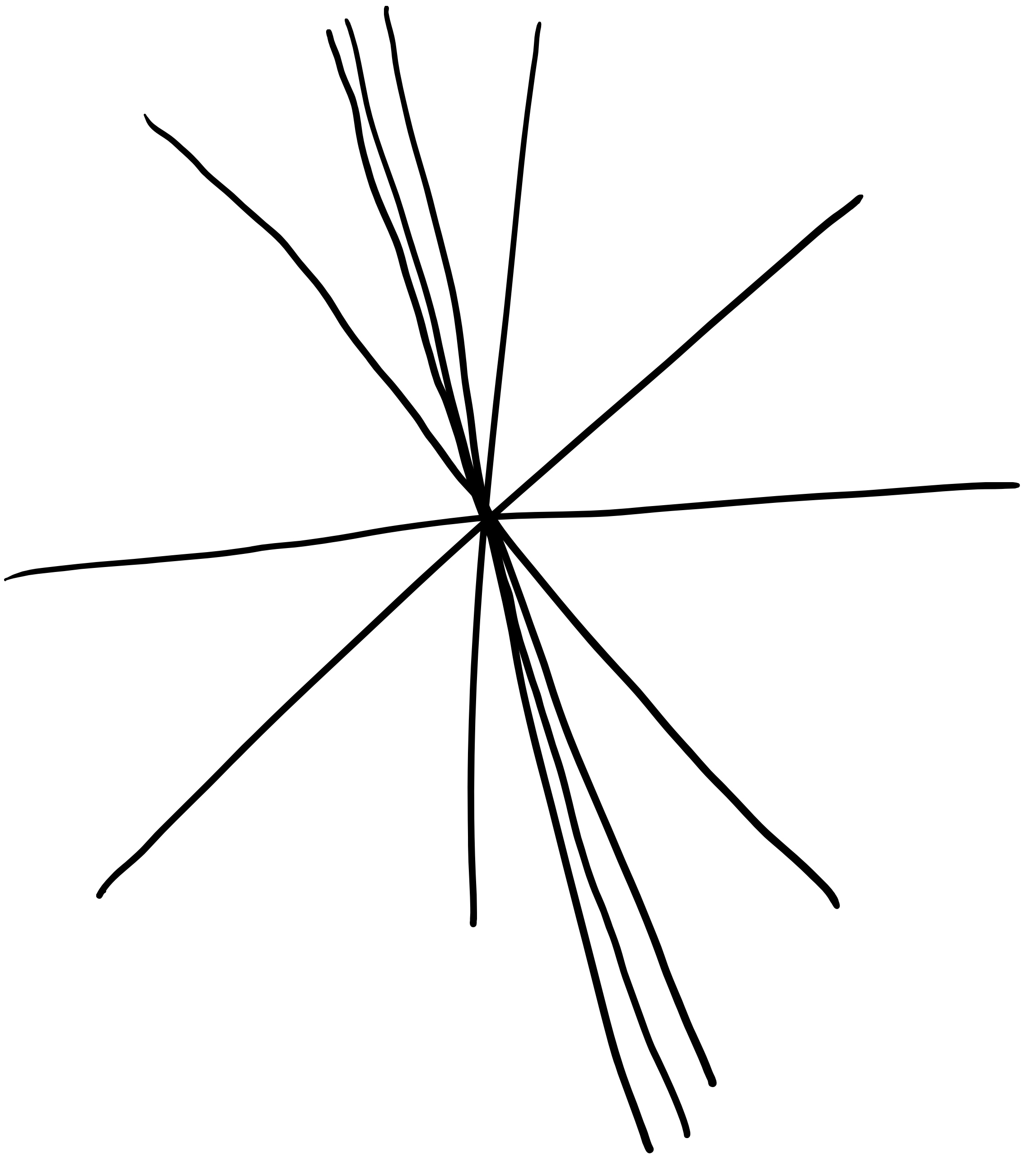
in L^p, for any $p \in [1, \infty)$.

Thm:

For $S = (N_0, N_1, 1)$ and ρ the ReLU, $\mathcal{R}(\mathcal{NN}(S))$ is closed in L^∞.

Conjecture:

Theorem holds for $L > 2$.



C. L. Frenzen, T. Sasao, and J. T. Butler. On the number of segments needed in a piecewise linear approximation. *Journal of Computational and Applied mathematics*, 234(2):437–446, 2010.

R. Eldan and O. Shamir. The power of depth for feedforward neural networks. In *Conference on learning theory*, pages 907–940, 2016.

T. Poggio, H. Mhaskar, L. Rosasco, B. Miranda, and Q. Liao. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *Int. J. Autom. Comput.*, 14(5):503–519, 2017.

U. Shaham, A. Cloninger, and R. R. Coifman. Provable approximation properties for deep neural networks. *Appl. Comput. Harmon. Anal.*, 44(3):537–557, 2018.

C. K. Chui and H. N. Mhaskar. Deep nets for local manifold learning. *Frontiers in Applied Mathematics and Statistics*, 4:12, 2018.

A. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inf. Theory*, 39(3):930–945, 1993.

Gitta Kutyniok, Philipp Petersen, Mones Raslan, and Reinhold Schneider, *A theoretical analysis of deep neural networks and parametric PDEs*, 2019, arXiv preprint arXiv:1904.00377.

Fabian Laakmann and Philipp Petersen, *Efficient approximation of solutions of parametric linear transport equations by ReLU DNNs*, Advances in Computational Mathematics **47** (2021), no. 1, 1–32.

Andrei Caragea, Philipp Petersen, and Felix Voigtlaender, *Neural network approximation and estimation of classifiers with classification boundary in a Barron class*, 2020, arXiv preprint arXiv:2011.09363.

Philipp Petersen, Mones Raslan, and Felix Voigtlaender, *Topological properties of the set of functions generated by neural networks of fixed size*, Foundations of Computational Mathematics (2020), 1–70.

Thank you for the attention!